SYSTEM, METHOD, AND COMPUTER PROGRAM PRODUCT FOR IDENTIFYING MULTI-PAGE DOCUMENTS IN HYPERTEXT COLLECTIONS

Abstract of the Disclosure

A system, method, and computer program product for identifying compound documents as a coherent body of hyperlinked material on a single topic as created by an author or collaborating authors, analyzing the content and structure of the compound documents and related hyperlinks, and responsively selecting a preferred entry point at which to begin processing such documents. The body of material may include the internet, an intranet, or other digital library that typically has content distributed over several separate pages or URLs, sometimes in a hierarchical directory structure. The processing may include creating at least one taxonomy, as well as searching or indexing the compound documents. The identification and analysis schemes include a observation of a number of heuristics run on component documents in the compound documents.